The 2nd Challenge on Remote Physiological Signal Sensing (RePSS)

Xiaobai Li[1], Haomiao Sun[2], Zhaodong Sun[1], Hu Han[23],
Antitza Dantcheva[4], Shiguang Shan[2], Guoying Zhao[1]
[1]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, China
[3]Peng Cheng Laboratory, Shenzhen, China.
[4]STARS team, INRIA, France
{xiaobai.li, zhaodong.sun, guoying.zhao}@oulu.fi,
haomiao.sun@vipl.ict.ac.cn
{hanhu, sgshan}@ict.ac.cn, antitza.dantcheva@inria.fr

Abstract

Remote measurement of physiological signals from videos is an emerging topic. The topic draws great interest, but the lack of publicly available benchmark databases and a fair validation platform are hindering its further development. The RePSS Challenge is organized as an annual event for this concern. Here the 2nd RePSS is organized in conjunction with ICCV 2021. The 2nd RePSS contains two competition tracks. Track 1 is to measure inter-beat-intervals (IBI) from facial videos, which requires accurate measurement of each individual pulse peak. Track 2 is about respiration measurement from facial videos, as respiration is another important physiological index related to both health and emotional status. One new dataset is built and shared for Track 2. This paper presents an overview of the challenge, including data, protocol, results, and discussion. We highlighted the top-ranked solutions to provide insight for researchers, and we also outline future directions for this topic and this challenge.

1. Introduction

Physiological signals such as the heart rate (HR), respiration rate (RR), and heart rate variability (HRV) are important indicators of human physical conditions. Traditional ways of measuring physiological signals always involve using special medical instruments such as the electrocardiography (ECG), the Photoplethysmograph (PPG) oximeter, the breathing belt and so on. Measurement with contact medical sensors is costly and not convenient especially for long-term use. More than ten years ago, researchers found that PPG signals can be remotely captured from human faces under ambient light. For example, Verkruysse et al. [20] reported measuring PPG signals from the forehead area. Many works followed after proposing various remote PPG (i.e., rPPG for short) measurement approaches. Early proposed methods are mostly unsupervised, which relied on filters designed with empirical knowledge and did not involve training process. Some of the approaches [15, 16, 3, 8, 11, 22] used the subtle color changes of the facial pixels for rPPG measurement, while some others [1, 12, 24] tracked vertical head motions for the task. Later more researchers started to exploit supervised approaches (e.g., [25], [26], [2], and [19]) for rPPG measurement. More details of the development of remote HR measurement are referred to survey papers [17] and [18].

Despite the thriving research interests, the lack of publicly available benchmark databases and a fair validation platform are the major issues that hinder its further development. For this concern, we organized the 1st RePSS challenge [1] in conjunction with CVPR 2020, aiming to provide benchmark datasets and a fair comparison platform for researchers working on this topic. The RePSS challenge was planned as a challenge series which we hope to organize one time per year with a continuous and proceeding theme. As the starting one, the 1st RePSS focused on the most fundamental task of measuring the average HR from color facial videos. This time, the 2nd RePSS is held in conjunction with ICCV 2021, which takes one step further, containing two challenge tracks and adding one newly built dataset.

The rest part of the paper is organized as follows: Section 2 gives an overview of the 2nd RePSS challenge, including the tasks of the two tracks, datasets, challenge protocol, and

[1] https://competitions.codalab.org/competitions/22287

evaluation metrics; Section 3 introduces some approaches proposed by participating teams that achieved leading performance in the challenge, Section 4 summarizes challenge results and discussions, and Section 5 discusses future directions of this topic.

2. Challenge Overview

2.1. Challenge tracks

There are two parallel tracks of the 2nd RePSS.

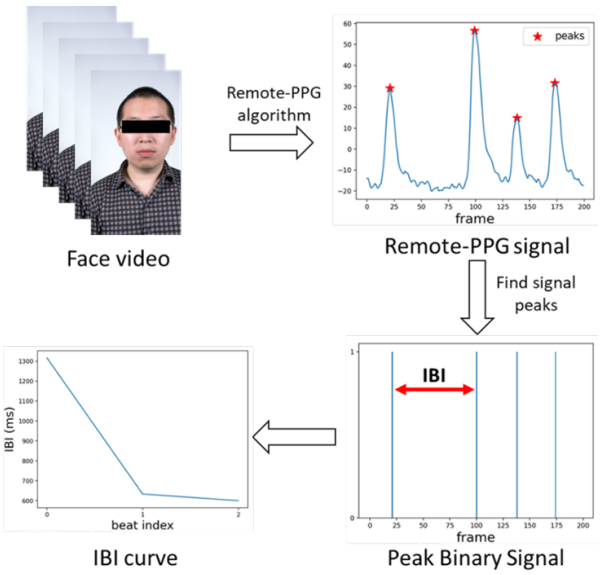

Figure 1. An IBI curve and peak binary signal from a face video.

Track1 requires measuring *inter-beat-interval (IBI)* from facial videos. Previous studies only focused on measuring the average HR. But in many applications, the average HR is not enough and more detailed information such as HRV features are needed. The IBI curve requires accurate measurement of each heartbeat, which can be processed to calculate various HRV features. Figure 1 shows how a peak binary signal and an IBI curve are obtained from a facial video. We first locate the peaks from rPPG signals and calculate the time interval between every two peaks. The red arrow shows one time interval. Thus, we can plot the IBI curve with the x-axis as the beat index and the y-axis as the IBI values.

Track2 requires measuring *respiration rate (RR)* from facial videos. The teams can use any clue or approach, e.g., color, motion, frequency domain analysis, etc., to estimate the respiration frequency from facial videos. Face videos and corresponding breathing curves measured with a breathing belt will be provided for training.

2.2. Data

The data used for the 2nd RePSS challenge come from three databases. Besides the VIPL-HR-V2 [13, 14] and the OBF [7] which were used in the 1st RePSS, we built and added a new dataset RePSS-RESP for the Track2 of respiration frequency measurement. Descriptions about the OBF and the VIPL-HR-V2 can be found in the previous RePSS challenge paper [9] or the original database papers.

RePSS-RESP is a new dataset collaboratively built by the Institute of Computing Technology of Chinese Academy of Sciences and the Center for Machine Vision and Signal Analysis of University of Oulu. RePSS-RESP contains upper body videos and corresponding breathing signals recorded from ten subjects while exercising in a gym. One main motivation of building the RePSS-RESP dataset is to provide data for Track2, and we involve two exercises at different intensity levels to include a wider span of respiration rates. The ten subjects are all Asians (mean age = 24), of which nine are males. Each subject performed two kinds of exercises, i.e., running on a treadmill, and cycling on a spinning bike, each for about 15 minutes for a total duration of approximately 30 minutes. The exercises programs were pre-set on the devices with varying intensity to achieve a wider range of respiration frequency. A Logitech camera (c1000e) was used for recording videos with a frame rate of 30 frames per second (fps) and a resolution of 1920 by 1080. A Go Direct respiration belt [2] was used for recording breathing signals as the ground truth at a sampling rate of 10 Hz. Before the data collection, every subject was informed about the research purpose and the recording procedure, and a consent form was signed if all conditions were clearly understood and accepted. The respiration belt was tied around the subject's chest and remotely connected to a laptop via Bluetooth, and the camera was fixed on a tripod and placed about one meter away in front of the subject. Sample frames from the RePSS-RESP videos are shown in Figure2.

Track1 training data The training data of Track1 is the same as used in the 1st RePSS challenge. 500 subjects' data from the VIPL-HR-V2 database are used for training. For each subject, we randomly selected five ten-second clips, which add up to altogether 2500 video clips. The videos are at an average speed of 25 fps with the resolution of 960 by 720. One thing to mention is that the frame rates of the training videos vary slightly that need to be concerned and countered by participating teams. Since the task is to measure IBI, we provide the corresponding PPG signal curves as the ground truth for training.

Track1 testing data The testing data of Track1 is the same as used in the 1st RePSS challenge so that the new

[2] https://www.inds.co.uk/product/go-direct-respiration-belt/

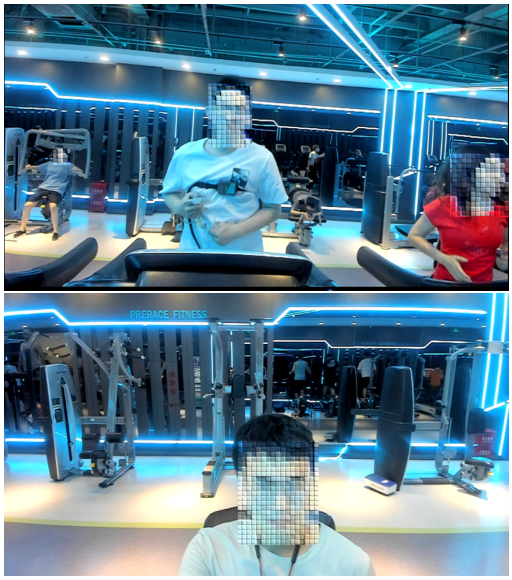

Figure 2. Sample frames from the RePSS-RESP dataset. Upper: running; lower: cycling.

results can be compared with those of the 1st RePSS challenge. The testing data consists of two parts, i.e., 100 subjects (no overlap with the training set) from the VIPL-HR-V2 database and 100 subjects from the OBF database. For each subject there are five ten-second clips, which add up to altogether 1000 samples. The OBF videos have a fixed frame rate of 30 fps with the resolution of 1920 by 1080. All the test videos were anonymized by covering facial features with mosaic blocks to protect identity information. Face positions and facial landmark locations were provided [3] to facilitate the testing process if needed. Ground truth IBI curves were computed from corresponding BVP signals of both databases, which were not provided to challenge participants and only be used for the evaluation carried out the challenge organizers based on the results submitted from the participants.

Track2 training data The training data are from the OBF dataset, which contains 100 subjects' face videos with a resolution of 1920*1080 at 30 fps. For each subject, there are ten 60-second clips. The corresponding respiration signals measured with a breathing belt at 256 Hz were provided as the ground truth signals for training.

Track2 testing data The testing data of Track2 contains the whole data of the newly recorded RePSS-RESP dataset. The 30-minute video (15 minutes of cycling and 15 minutes of running, some recordings are slightly shorter) of each subject is divided into 60-second short clips as the testing samples, ending up to altogether 283 testing samples (134 running clips and 149 cycling clips). The average RR computed from the ground truth respiration signals are not provided but used for evaluating the submitted results.

2.3. Challenge protocol

The 2nd RePSS challenge is operated on the CodaLab platform [4], consisting two stages as follows.

Training phase (28.04.2021 – 10.06.2021) The 2nd RePSS opened the challenge on the 28th April 2021 on the CodaLab website, and the tasks of the two tracks were announced. The training data of both tracks were released on the 15th. May 2021. Participants can register either as an individual or as a team to the challenge. The registration closed by the 10th June, 2021. We require that one registered team can have up to ten team members. The participating teams have to sign license agreements before achieving the data for the challenge. No result submission could be made to the challenge website during the training phase.

Testing phase (25.06.2021 – 12.07.2021) The testing data of both tracks were released on the 25th. June 2021. All participating teams can submit their testing results to the website from the date of testing data release until the 9th June 2021. As observed from the 1st RePSS challenge of the last year, some participants registered multiple CodaLab IDs in order to submit more rounds of test results. To improve the challenge in a fairer way, this time we require that every participating team can only submit results under a pre-authorized ID name and results submitted under other unauthorized CodaLab IDs won't be listed in the final ranking. For each ID, there can be maximum 30 times of submissions. The best performance of all submissions from each team will be used for the final ranking.

2.4. Evaluation metrics

For Track1, one peak binary signal (as shown in the lower right part of Figure 1) should be submitted for one test video which contains only zeros and ones. The ones correspond to the heartbeat peaks and the zeros for no peak. The length of the binary signal should be equal to the frame number of the video, i.e., for a video of 300 frames, the peak binary signal should be a 300-dimensional binary vector.

The submitted peak binary signals are used to compute five metrics for evaluation and ranking, of which three metrics are on HR level and two metrics are on IBI level: 1). Mean absolute error of heart rate: MAE_HR. 2). Root mean squared error of heart rate: $RMSE\_HR$. 3). Pearson correlation coefficient of heart rate: R_HR. 4). The mean of IBI error: M_IBI. For two IBI curves $R_1(t)$ and $R_2(t)$, the M_IBI can be defined as,

$$AE = \sum_{t=0}^{T} |R_1(t) - R_2(t)|, M\_IBI = \frac{\sum_{k=0}^{K} AE_k}{K}, \quad (1)$$

[3] https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format

[4] https://competitions.codalab.org/competitions/30855

Table 1. The final result leaderboard of the 2nd challenge of RePSS.
(a) Track 1

Team Name	Organization	IBI		HR		
		M	SD	MAE	RMSE	R
TIME	Shandong University	**117.25**	**153.18**	7.31	11.44	**0.62**
Dr. L	Hefei University of Technology	122.80	153.91	**7.29**	**11.05**	0.57
The Anti-Spoofers	Tencent YouTu Lab	168.08	162.82	11.84	14.51	0.02
shankejinjiboy	Shandong University of science and technology	224.41	163.98	15.44	18.75	-0.05
ZJUT-WTCrPPG	Zhejiang University of Technology	273.53	171.13	23.89	27.96	-0.03
ZJUT-ASTrPPG	Zhejiang University of Technology	295.70	175.24	29.24	33.69	-0.10

(b) Track 2

Team Name	Organization	MAE_RR	$RMSE\_RR$	R_RR
PCALab_DeepInsight	Nanjing University Of Science And Technology	**7.22(1)**	**8.85(1)**	0.01(2)
JDD	Hong Kong Baptist University	7.56(2)	9.04(2)	0.01(3)
Dr. L	Hefei University of Technology	7.58(3)	9.60(3)	**0.44(1)**

where T is the time length of the IBI curves, and K is the number of videos. 5). The standard deviation of IBI error: SD_IBI, which can be written as,

$$SD\_IBI = \sqrt{\sum_{k=0}^{K}(AE_k - M\_IBI)^2/K} \qquad (2)$$

The first three metrics are the same as used in the 1st RePSS challenge, and the last two are new for evaluating model performance on the IBI level. The M_IBI will be used as the main metric for ranking, and other metrics will also be shown for reference.

For Track2, one average respiration frequency, e.g., 30 breaths per minute (bpm), should be submitted for one testing video. Three metrics are calculated for evaluation, including: 1). the mean absolute error of respiration rate: MAE_RR, 2). the root mean squared error of respiration rate: $RMSE\_RR$, and 3). the Pearson correlation coefficient of respiration rate: R_RR. The MAE_RR will be used as the main metric for ranking, while other metrics will also be shown for reference.

3. Proposed approaches

To facilitate a fair competition, only pre-registered teams with authorized team IDs are included in the final ranking. There are altogether 18 teams registered for the challenge, from which 17 teams registered for Track1 and ten teams registered for Track2 (nine teams registered for both tracks). Till the final submission date, there were valid results submissions of six teams for Track1 and three teams for Track2 (one team submitted for both tracks) which are included in the final ranking. The teams info and their performances are shown in Table 1. We contacted all teams to provide a brief description of their methods to be included in this review paper. Five teams approved the context usage and their methods are introduced in the following.

3.1. Team 'The Anti-Spoofers' (Tencent YouTu Lab)

The team Anti-Spoofers participated in Track1 of IBI measurement, and they proposed a CNN based framework for measuring IBIs from facial videos. The overview of The Anti-Spoofers' method is shown in Figure 3. The method contains three steps: 1). data pre-processing, 2). BVP signal estimation via an end-to-end efficient framework, and 3). post-processing of the predicted signals.

In data pre-processing, the left and right cheeks are chosen as the ROI regions and transformed to a square as the input basing on the facial landmarks of each video frame. Then, the ground truth BVP signals are filtered by a bandpass filter to smooth the curves and facilitate the training. As for the heartbeats, we calculate them via heart rate as an auxiliary supervision to the framework.

To predict BVP signals, we propose an end-to-end efficient framework, which includes a Feature Extractor, a BVP Estimation Network and a Heartbeat Measuring Network. Specifically, the Feature Extractor is composed of spatial convolutions and temporal convolutions to extract rPPG related features from the input with low cost of calculation. Then, BVP Estimation Network consists of 1D convolutions to estimate BVP signals based on extracted features at frame level. To facilitate the learning of the BVP Estimation Network, a Heartbeat Measuring Network is also introduced which has a similar structure as the BVP Estimation Network. It predicts heartbeats at video level by compressing the information in the temporal dimension as an auxiliary supervision.

In the post-processing step, the estimated signals are filtered with a bandpass filter and peaks are detected by com-

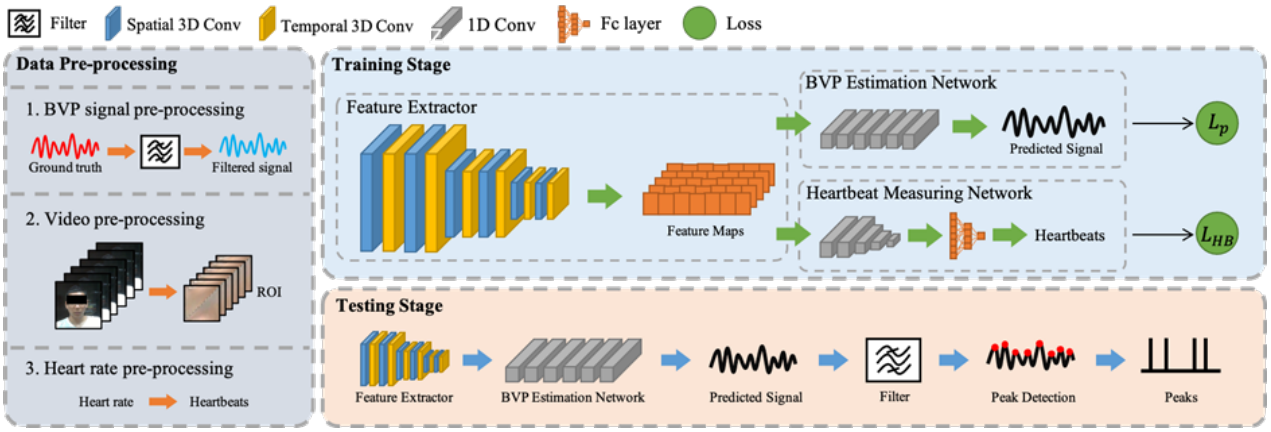

Figure 3. Overview of The Anti-Spoofers' method.

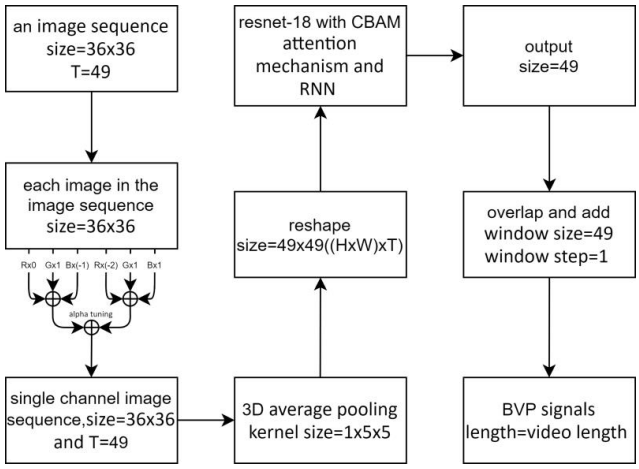

Figure 4. Overview of team shankejinjiboy's method.

puting the first order difference of the signals to output peak binary signals as requested by the challenge.

3.2. Team 'shankejinjiboy' (Shandong University of Science and Technology)

The team shankejinjiboy participated in Track1 of IBI measurement, and they proposed a neural network with attention mechanism for the task. Firstly, the face parts detection toolbox of MATLAB [5] is used to detect faces. After detecting the faces, the eyes and background are removed. Then, our method is mainly composed of the following three steps as represented in Figure 4.

Reserved space context information: The POS algorithm [22] is used to process the images with space context information. The input is a 36x36 image sequence with a time length of T = 49. For each image of an image sequence, the spatial context information is reserved, and then the three color channels are multiplied by the corresponding weights respectively. The corresponding weight matrix is $\begin{pmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \end{pmatrix}$. Then the three channels are summed up as one channel. Because there are two different weights of RGB channels, two channels are obtained. These two channels pass through an alpha tuning to obtain a single channel image. Finally, the output of this processing is a single channel 36x36 image sequence with T = 49.

Neural networks and attention mechanism: For a single channel image sequence with T = 49 and size = 36x36, one 3D average pooling layer with a kernel size of 1x5x5 is used for processing in order to reduce the quantization error of the camera. Then, the pixels of each image in the image sequence are arranged into a column vector according to the size of H x W = 7x7 and T = 1. Finally, a single channel image with the size of H x W = 7x7 and T = 49 is obtained. This image is used as the input and fed into an Restnet-18 proposed in [5], which uses the CBAM attention mechanism [23]. The next step is to replace the spatial attention mechanism in CBAM with a spatial attention mechanism based on channel attention mechanism. This attention mechanism is used to find the average value on the T dimension, which compresses the image into a column vector of H and T = 1. Then two 1D convolution layers with a convolution kernel of 1 are used to obtain one single channel column vector of H and T = 1. Multiply this column vector by the column vector of each column of the image before processing to obtain an image with attention weight. The final output of the Restnet-18 is a vector of length of 49, which corresponds to the BVP signal value of T = 49. Finally, a RNN is used to fit this vector and the real BVP signal value, which is to eliminate the delay errors and motion noises between estimated face rPPG signals and ground truth PPG signals.

Overlap and generate BVP signal: We need a smoothing process to get a smooth BVP signal. For the image sequence input to the neural network in the previous step, the

[5] https://www.mathworks.com/matlabcentral/fileexchange/36855-face-parts-detection

image sequence of T = 49 and size = 36x36 is obtained by using a window length of T = 49 frames and a step size of 1 frame. Similarly, the output of the neural network in the previous step is a BVP signal with the same length of T = 49 and a step size of 1. The output BVP signals are overlapped and added to obtain a smooth BVP signal.

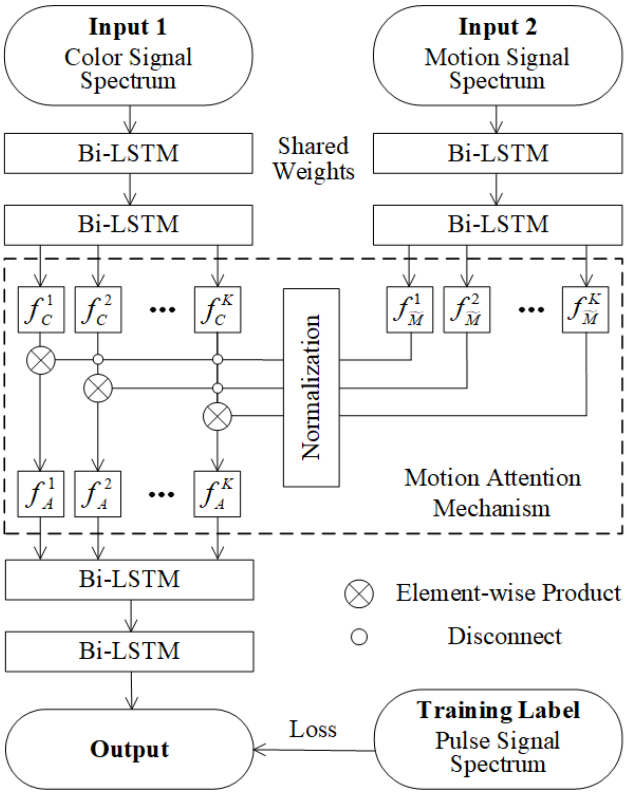

Figure 5. Overview of team Dr.L's method.

3.3. Team 'Dr.L' (Hefei University of Technology)

The team Dr.L participated in Track1 of IBI measurement, and they proposed a method which fuses both color and motion information for measuring IBIs from facial videos. The framework of the proposed method is shown in Figure 5.

First, a skin color signal and a nose moving signal are obtained from each facial video, and the frequency spectrum of the two signals are computed. The motion-driven attention network (MANet) takes the frequency spectrum of a skin color signal and a nose moving signal as the inputs, and extracts their features using bidirectional LSTM networks. Although the motion signal mixed in the skin color signal differs from the nose moving signal in waveform, they are similar in the distribution of the frequency components. Thus, the spectral features of the nose moving signal are used as an attention mask covering on the spectral features of the skin color signal, with the purpose of suppressing the motion components in the spectral features of the skin color signal. The features covered with the mask are then fed into bidirectional LSTM networks to reconstruct the pulse signal spectrum, which is finally transformed back to a pulse signal.

3.4. Team 'TIME' (Shandong University)

The team TIME participated in Track1 of IBI measurement, and they used an rPPGNet-based approach with data augmentation for measuring IBIs from facial videos. The framework is shown in Figure 6 which contains three stages with data augmentation.

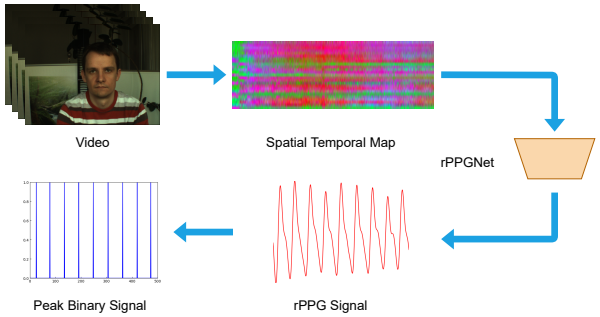

Figure 6. Overview of team TIME's method.

Data augmentation: Four data enhancement strategies are adopted to make the deep learning model more robust. (1) randomly erase part of STmap; (2) randomly add random noise to part of STmap; (3) randomly reverse STmap and the corresponding ground-truth rPPG signal; (4) randomly flip facial video horizontally.

Pre-processing: First, the method detects faces using RetinaFace[4] with MobileNet backbone[6]. Then, face alignment is performed for each face using the eye centre points. Before generating STmap, all face videos and the corresponding rPPG signals are re-sampled to 30 fps using cubic spline interpolation like [10]. Finally, the ground-truth rPPG signal is filtered using a 4th-order Butterworth band-pass filter with cutoff frequency [0.6, 3] Hz like [21] and normalized to have a minimum value of zero and a maximum value of 1.

rPPG signal estimation neural network: A convolutional neural network model named rPPGNet is proposed for rPPG signal estimation. The input data of rPPGNet is STmap like [14]. Three external datasets are used for training(VIPL-HR, PURE, UBFC-rPPG).

Post-processing: For restricting outliers, the estimated rPPG signal is filtered through a 4th-order Butterworth band-pass filter with cutoff frequency [0.6, 3] Hz. After that, $scipy.signal.find\_peaks$ is used to find peaks of rPPG signal.

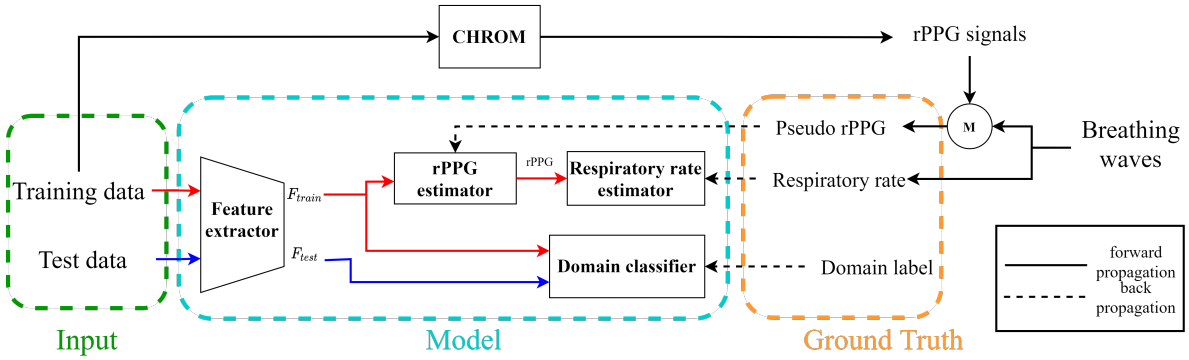

Figure 7. Overview of team JDD's method.

3.5. Team 'JDD' (Hong Kong Baptist University)

The team JDD participated in Track2 of respiration measurement, and they proposed a weakly supervised network for measuring respiration frequency from facial videos. The framework of the proposed method is shown in Figure 7.

The JDD researchers proposed a novel weakly-supervised domain adaptive end-to-end network for respiratory rate estimation. The network estimates both the rPPG signal from input and analyzes the respiratory pattern in the rPPG signal for respiratory rate estimation. We generated the pseudo label for rPPG estimation by modulating the CHROM [3] estimated rPPG with the ground truth breathing wave. To overcome the domain shift between the training dataset and the test dataset, a customized domain adaptation is applied on augmented training and test dataset.

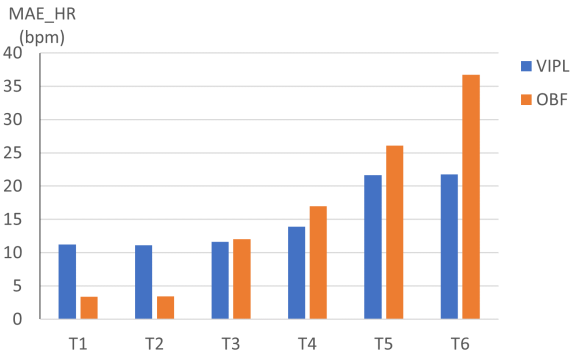

Figure 8. Compare the performance on VIPL-HR-V2 and OBF separately. Lower values of MAE_HR indicate better performance. T1, T2 ... T6 indicate participating teams.

4. Challenge results and discussion

The main results are summarized and shown in the Table 1. In this section we also provide more detail statistics of the results, e.g., on separate datasets, for both tracks. The teams will be referred as T1, T2, ..., as their corresponding rank orders for convenience. For example, team 'JDD' will be mentioned as Track2 T2 for discussion.

4.1. Track1 results analysis

Track1 uses the same training and testing datasets as used in the 1st RePSS challenge, as we hope the results of the two rounds can be compared. The distribution patterns of the training and testing samples can be found in the 1st RePSS review paper [7].

First, we compare the performance on the average HR level using the metric of HR_MAE. As the testing set contains videos from VIPL-HR-V2 and OBF, we compare the mean absolute HR errors of the top three teams separately on the two datasets. The performance of the six teams in Track1 is shown in Figure 8. It can be seen from the figure that T1 and T2 performed significantly better on the OBF data than the VIPL-HR-V2, while for the other teams it was the otherwise. Considering the video quality, the OBF dataset should be easier than the VIPL-HR-V2 dataset, as the OBF videos have higher resolution, larger face size, and contain less subject motions. But on the other side, as the training data are all from VIPL-HR-V2, there might be domain transfer issue when testing on OBF samples, which is depending on the designed models. One possibility is that the approaches of T3 to T6 might be more impacted by the domain shift issue than those of T1 and T2. Besides, if compared with the performance of the 1st RePSS challenge, the top three teams of this year achieved a similar level of accuracy than the top three teams of the 1st RePSS concerning the mean absolute error of the average HR.

Second, we also compare the performance on the IBI level on the two datasets separately. The two metrics of M_IBI and SD_IBI are used for the comparison, and lower values indicate higher accuracy for both metrics. The performance of the six teams in Track1 is shown in Table 2. It can be seen that performance on the IBI level shows a similar trend as the average HR level. For the top two teams T1 and T2, their accuracy on the OBF dataset are significantly higher than on the VIPL-HR-V2 dataset, while for the other

four teams, it is the opposite. T3 accuracy achieved a similar level of performance as the top two teams on the VIPL-HR-V2 dataset, while their performance on the OBF dataset is significantly lower, which might be caused by the domain shift problem.

Table 2. Compare IBI measurement on VIPL-HR-V2 vs. OBF. Both metrics of M_IBI and SD_IBI are in millisecond (ms), and lower values indicate better performance. T1, T2 ... T6 indicate Track1 participating teams.

	VIPL-HR-V2		OBF	
	M_IBI	SD_IBI	M_IBI	SD_IBI
T1	150.15	159.90	84.34	138.55
T2	160.76	155.69	84.83	142.31
T3	164.67	159.29	171.49	166.20
T4	228.57	163.76	220.25	164.1
T5	264.41	168.77	282.64	172.97
T6	262.39	165.62	329.01	178.24

4.2. Track2 results analysis

It can be seen from the Table 1 that the gaps between all three teams of Track2 are very small. Comparing to the HR, the RR has a narrower distribution range. The Normal RR for an adult at the resting state is about 12 to 20 bpm, which could rise to 40 to 50 bpm when doing exercise depending on the intensity.

The testing data includes two kinds of exercises, i.e., running and cycling, so we split the data and compare the performance on the two parts. All three metrics of MAE, SD, and R are calculated separately on the two exercises' subsets, and the results are shown in Table 3. Generally speaking, the RR measurement seems to be more challenging under 'Running' than 'Cycling' as all three teams achieved smaller MAE and SD on 'Cycling'. One reason might be that cycling involves fewer upper body motions (or of lower intensity) than running. Besides, one thing to mention is that T3 achieved significantly better correlation R values than the other two teams, while their MAEs and SDs are at a similar level.

Table 3. Compare performance on 'Running' vs. 'Cycling' clips. Lower values of MAE and SD and larger values of R indicate better performance. T1, T2 and T3 indicate participating teams.

	Running			Cycling		
	MAE	SD	R	MAE	SD	R
T1	8.39	10.24	-0.03	6.19	7.38	0.11
T2	8.42	10.04	0.09	6.79	8.04	-0.23
T3	8.52	10.72	0.47	6.73	8.47	0.19

The results indicate that the current data is very challenging for all participating teams. Firstly,there are large differences between the training and testing data. We map the distribution patterns of the training data and testing data in Figure 9. It can be seen that there is distribution difference between the training and testing data. The training data (grey) are from OBF, which has a narrower RR range with most samples gathering at 12 to 25 bpm (mean = 17.53, Std = 3.96). The testing data are from the RePSS-RESP (blue) containing running and cycling clips with a wider RR range spreading from 10 to 40 bpm (mean = 25.02, Std = 8.59). The distribution differences may cause domain shift issue for the models. Secondly, all test videos were recorded in a gym under exercise mode, the environmental lighting and subjects' motions are more challenging than those of the training data which were recorded in a lab with subjects in a sitting position.

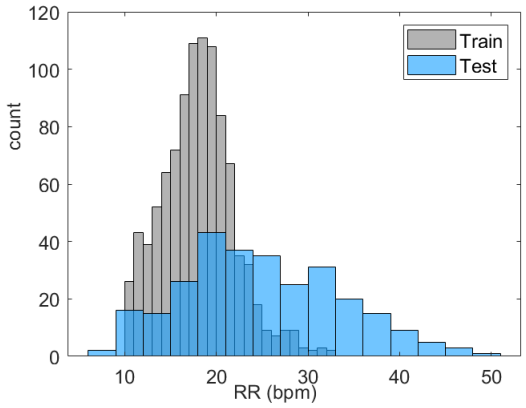

Figure 9. RR distributions of training vs. testing data.

Several factors were considered when we designed the scenario setups for collecting the RePSS-RESP dataset. First, we hope the data could include various human status but not just sitting. Second, we hope to cover a wider span of the RR or the samples points might just all cluster in a narrow range which is not efficient for developing RR measurement approaches in the long run. The current Track2 is very challenging mostly due to the big difference between training and testing data. In the future the new and old datasets will be combined and better balanced for training and testing.

5. Conclusion and future directions

As a continuous event, the 2nd RePSS challenge took one step further than the 1st RePSS from two aspects: 1). Track1 increased the task difficulty from 'measuring the average HR' to 'measuring individual heartbeats and the IBIs' as the Track1, and 2). added one new modality of measuring the respiration rate as the Track2. More focused research groups were involved, and interesting approaches were proposed from participating teams which might bring insights for future studies.

The Track1 task requires the proposed models to output rPPG curves instead of just one single value as the average HR in order to compute the binary IBI signals. Most of the proposed framework involved deep learning models and some pre-processing and post-processing steps. Based on observations of the results, domain shift issue might be one problem at least for some proposed methods, as the performance dropped on the novel test set (the OBF dataset for Track1 testing), and the Top one performing team reported using extra training data with data augmentation.

The Track2 task is an exploration of a new modality of breathing. Although the RR measurement was reported in several previous studies, it was less attended than the measurement of HR. A new dataset, the RePSS-RESP was built and shared for Track2, which contains about 300 minutes of data from ten subjects performing running or cycling exercises. The results performance showed that the task was very challenging for all participating teams, probably due to that 1). big distribution difference between the testing and training data, and 2) large motions and complex environment are involved in the testing data.

In the future, we plan to continue the RePPS challenge from the following aspects. First, the main task of HR and HRV (IBI) measurement will continue, but may be extended with extra data to explore more about the domain shift issue, or be specialized and split into several sub-tasks, e.g., to evaluate the pre-process, the main model, or the post-process step separately. Second, the respiration modality may also continue as respiration is very important in many applications. We will balance the training and testing datasets to adjust the task level. Third, in the future challenges, we would also like to consider a multimodal fusion approach for the rPPG measurement, e.g., to provide multiple sources of simultaneously recorded videos that could be fused for more robust rPPG measurement.

6. Acknowledgement

We would like to thank the ICCV 2021 conference organizers for agreeing to host our challenge. Xiaobai Li and Guoying Zhao's work were supported by National Natural Science Foundation of China (Grant 61772419), and Academy of Finland (Grant 316765 and 323287). Hu Han's work was supported in part by Natural Science Foundation of China (Grant 61672496).

References

[1] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proc. IEEE CVPR*, pages 3430–3437, 2013. 1

[2] Weixuan Chen and Daniel Mcduff. Deepphys: Video-based physiological measurement using convolutional attention networks. *Proc. ECCV*, pages 356–373, 2018. 1

[3] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Trans. Biomed. Eng.*, 60(10):2878–2886, 2013. 1, 7

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 6

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, pages 770–778, 2016. 5

[6] Zhu M. Chen B. Kalenichenko D. Wang W. Weyand T. Adam H. Howard, A. G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv*, 2017. 6

[7] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *Proc. IEEE FG*, pages 1–6, 2018. 2, 7

[8] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proc. IEEE CVPR*, pages 4264–4271, 2014. 1

[9] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[10] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12404–12413, 2021. 6

[11] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Trans. Biomed. Eng.*, 61(10):2593–2601, 2014. 1

[12] Andreia Vieira Moc¸o-, Stuijk Sander, and Gerard de Haan. Ballistocardiographic artifacts in ppg imaging. *IEEE Trans. Biomed. Eng.*, 63(9), 2016. 1

[13] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Proc. ACCV*, pages 562–576, 2018. 2

[14] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. Image Processing*, 2019. 2, 6

[15] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, 2010. 1

[16] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological

measurements using a webcam. *IEEE Trans. Biomed. Eng.*, 58(1):7–11, 2011. 1

[17] Philipp V. ROUAST, Marc T.P. ADAM, Raymond CHIONG, David CORNFORTH, and Ewa LUX. Remote heart rate measurement using low-cost rgb face video: a technical literature review. *Frontiers of Computer Science (electronic)*, 2018. 1

[18] Alexander Trumpp Daniel Wedekind Sebastian, Zaunseder and Malberg Hagen. Cardiovascular assessment by imaging photoplethysmography - a review. In *Biomed. Eng. - Biomed. Tech.*, 2018. 1

[19] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021. 1

[20] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, 2008. 1

[21] Wenjin Wang, Albertus C Den Brinker, and Gerard De Haan. Single-element remote-ppg. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2018. 6

[22] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Trans. Biomed. Eng.*, 64(7):1479–1491, 2017. 1, 5

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[24] Cheng Yang, Gene Cheung, and Vladimir Stankovic. Estimating heart rate and rhythm via 3D motion tracking in depth video. *IEEE Trans. Multimedia*, 19(7):1625–1636, 2017. 1

[25] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *Proc. BMVC*, 2019. 1

[26] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. In *Proc. IEEE ICCV*, October 2019. 1